

ASSESSMENT & INFORMATION

NCS Pearson, Inc.
2510 North Dodge
Iowa City, Iowa
52241
USA

www.pearson.com

April 30, 2012

Mr. Ken Slentz
Deputy Commissioner, Office of P-12 Education
New York State Education Department
89 Washington Avenue
Albany, New York 12234

Dear Ken,

Pursuant to our discussion of Sunday, April 29, 2012, Pearson stands behind our work in New York as we do the work provided by our subcontractors. As such, we are committed to eliminating any gaps identified by the New York State Education Department between expectation and our performance in the spring of 2012. In this regard, you identified two global issues I want to address: performance of our translation subcontractor (Eriksen Translations Inc.) and quality of the work supporting the New York State Testing Program-Scoring Guides.

Translation

As you may recall, Eriksen Translations Inc., is our identified subcontractor performing translations in New York in partial fulfillment of our Grade 3-8 assessment contract. Eriksen Translations Inc. is also an identified Minority and Women Business and, as such, helps both Pearson to fulfill contractual requirements as well as the state to meet its goals.

This spring, several translation issues were identified ranging from the lack of a correct response option for some multiple choice items (and / or different response options than in the source English version of the assessments) to omitted words or phrases, typesetting / formatting errors, errors in vocabulary or the translation itself. All of these issues were introduced during translation and the subsequent typesetting of the translated versions of the tests. Pearson and Eriksen are already documenting these issues and we are taking further actions to better understand why and how they occurred. For example, Eriksen is performing a "root cause" analysis working in coordination with Pearson Organizational Quality to identify required corrective actions. It is premature at this juncture to determine changes in procedures and processes until the investigation into cause is completed. However, we have identified several options for consideration regarding process improvement:

- Enterprise scheduling. One area of concern for translations is the amount of lead time required to accommodate the translation process. As you may recall, Eriksen must translate five different languages (traditional-Chinese, Haitian-Creole, Korean, Russian, and Spanish). Furthermore, Eriksen uses both a forward and backward translation process where the base English language test is forward translated into the various target languages which are then back translated to English and compared against the English language source. Such an iterative process allows for the correction of various aspects of the translation. This process, however, requires that Eriksen start with the final English language test forms. That did not occur this year. This year, because of the compressed schedule, Eriksen started the translation process during one of the review stages of the English language assessments. Many changes were introduced into the English language assessments *after* Eriksen started, causing unanticipated rework and versioning control issues. Going forward we plan to include Eriksen’s translation needs explicitly in the enterprise schedule such that we can quantify the risks that schedule changes will have on Eriksen’s ability to follow their necessary work flow.
- Production process. While the root cause analysis is not yet complete, many of the types of issues and errors seen to date involved typesetting and/or proof reading errors than actual translation errors. For example, incorrectly changing “(a+6) and (a-3)” to “(a+6) and (a+6)” is a typesetting and / or proof reading error rather than a translation error. We can address such issues by providing support to Eriksen for desk top publishing, proof reading or typesetting or provide an additional round of independent proof reading. Furthermore, making simple changes in how EPS files (Encapsulated PostScript Files)—which are typically self contained files for the transfer and display of graphics and art—are exchanged between Pearson and Eriksen can minimize the chance that additional errors will be introduced into the process.
- Independent verification. Since Eriksen uses a forward and backward translation process, it would be advisable to add a third party independent translator to verify and document the decisions made to resolve inconsistencies between the two versions (i.e., the English source and the English version resulting from translating back to English from the target language). It is at this stage where varied judgment regarding correct vocabulary use could affect the quality of the translation. For example, while it will not be known for certain until the root-cause analysis is performed, the incorrect use of the word “median” instead of the correct use of the word “mean” might have resulted due to decisions made during this stage.

Regardless of the specific actions taken (as guided by the root cause analysis and in consultation with the NYSED) Pearson is ready to improve the procedures and ultimate quality of the translation process and outcomes and these suggestions represent our earliest thoughts.

Scoring Guides

The complete review of the scoring guides, while not documenting any significant errors or immediate action items, did reveal that the scoring guides in general need improvement to become the exemplar documents expected by the NYSED. Pearson agrees that we need to work diligently to improve these guides used by teachers to score constructed response items in New York. While we need more time to pull together a comprehensive plan (working with our own scoring experts and process engineers) some of our ideas for immediate action include:

- Mining information from field testing, prompt and rubric development. Typically, during the development of a constructed-response item, the logic for a fully correct score and each partially correct score is documented and translated into rules for scoring. While this is standard practice there are additional processes that can be undertaken. For example, during field testing, a host of unanticipated, but potentially correct (as well as incorrect) answers will be obtained from students. These answers are typically reviewed to verify that the anticipated correct answers are indeed discovered in student responses. In addition, we plan to review additional student responses for novel solution sets and document the various ways in which students obtain partial credit responses. Ultimately this might require a larger sample of student responses, but such data will allow us to document actual student performance across a variety of scenarios leading to potentially correct responses.
- Expert Review. Similar to the recent post-hoc review performed by the Regents Fellows and independent Pearson experts for the current scoring guides, we should also incorporate an expert-independent review into our process for the development of the scoring guides as routine process going forward. We are also considering hiring a dedicated scoring resource to work in and with the content development team to help align content and performance scoring activities.
- “Test Hacker” Review. One idea we have discussed that would be particularly applicable to constructed-response questions and their associated scoring rules would be to ask a team of savvy, subject matter experts who have not been associated with the item development to take the test with direction to find flaws, errors or otherwise defeat the assessment. We could then review the range of responses and/or interview these hackers to understand better what they tried and how robust items withstood various attacks. These same “hacker responses” can be scored using the developed scoring guides as another test of the ability of the scoring guides to provide correct partial and full credit responses.
- Expanded use of prototype items. Currently during field testing the prototype or exemplar items are chosen to represent a wider range of items regarding the development of scoring rules and guides. These items receive the full complement of anchor, practice, and qualification sets. We could expand this such that student responses and complete descriptions of the non-

prototype items also receive the full anchor, practice, and qualification sets. Currently the non-prototype items have only anchor and practice sets. Pearson could develop these complete training sets on all of the items generated and review in anticipation of developing a scoring guide for each and every item (even if we do not choose a particular item at that particular time to be included in an operational test).

Again, these are our immediate suggestions on how to improve the overall quality of the Scoring Guides. We would like to take additional time to develop and vet a more comprehensive plan with timelines, tasks, responsibilities and outcomes clearly articulated and documented.

Pearson is here to support you as we transition to cutting edge assessments measuring the Common Core State Standards and college and career readiness. As such, we strive for continuous improvement and pledge to continue to learn and improve as we work together. As always, if you have any questions, need clarification or additional information please drop me a note at jon.s.twing@pearson.com or call me at 319.331.6547.

Sincerely,

A handwritten signature in black ink, appearing to read "Jon S. Twing". The signature is fluid and cursive, with the first name "Jon" and last name "Twing" clearly distinguishable.

Jon S. Twing, Ph.D.

Executive Vice President & Chief Measurement Officer
Pearson